smalldisco, a pipeline for siRNA discovery and 3' tail identification

Ian V. Caldas[1,2], Leanne H. Kelley[2], Yasir H. Ahmed-Braimah[2*], and Eleanor M. Maine[2*]

[1]Department of Molecular Biology and Genetics, Cornell University, 526 Campus Rd, Ithaca, NY 14850
[2]Department of Biology, Syracuse University, 107 College Place, Syracuse, NY 13210

*Corresponding author: Email: yahmed@syr.edu (Y.H.A.); emmaine@syr.edu (E.M.M.)

Abstract

Capturing and sequencing small RNAs is standard practice, however identification of a group of these small RNAs—small interfering RNAs (siRNAs)—has been more difficult. We present smalldisco, a command-line tool for *small* interfering RNA *disco*very and annotation from small RNA-seq datasets. smalldisco can distinguish short reads that map antisense to an annotated genomic feature (e.g., exons or mRNAs), annotate these siRNAs, and quantify their abundance. smalldisco also uses the program Tailor to quantify 3' non-templated nucleotides of siRNAs or any small RNA species. smalldisco and supporting documentation are available for download from GitHub (https://github.com/ianvcaldas/smalldisco).

Short running head: smalldisco program for siRNA discovery

Keywords: small RNA, siRNA identification, small RNA 3' tailing modification, non-templated nucleotide addition

Introduction

Small, non-coding RNAs (sRNAs) are critical components of transcriptional and post-transcriptional gene regulatory mechanisms in plants, animals, and fungi (Castel and Martienssen 2013, Holoch and Moazed 2015, Ozata *et al.* 2019, Galagali and Kim 2020, Ketting and Cochella 2021). Three abundant types of sRNAs — micro (mi) RNAs, PIWI-associated (pi) RNAs, and small interfering (si) RNAs — associate with Argonaute proteins and provide specificity via complementary base pairing with target RNAs (Iwakawa and Tomari 2022). Although the details of sRNA biogenesis vary among species, distinct miRNA

and piRNA gene clusters are present in the genome, often under control of dedicated promoters, and expression of these genes produces populations of identical miRNA or piRNA sequences. In contrast, siRNAs in most organisms are generated by Dicer cleavage of double strand (ds) RNA, itself often generated by RNA-directed RNA polymerase (RdRP) activity on single strand (ss) RNA templates. One notable exception is nematodes, where most siRNAs are produced as short RdRP products that are not processed by Dicer. Complicating matters, biogenesis of different sRNA types can be linked, e.g., plants produce phased siRNAs downstream of miRNA-AGO activity on RNA pol II transcripts, and nematodes produce siRNAs downstream of piRNAs. Because miRNAs and piRNAs are identified as discrete, genome-encoded sequences, it is relatively straightforward to identify miRNA and piRNA sequences in sRNA-seq data sets by mapping reads to their canonical loci in the genome. In contrast, siRNA sequences often map to numerous positions across the template RNA and are not always straightforward to identify.

The main methods used for identifying siRNAs have included: (i) recovering sRNAs that co-immunoprecipitate with a specific Argonaute protein (e.g., Claycomb *et al.* 2009, Chaves *et al.* 2021, de Albuquerque *et al.* 2015; Mi *et al.* 2008; Seroussi *et al.*, 2022; Svendsen *et al.* 2019, Xu *et al.*, 2018; Okamura *et al.*, 2013; Gartland *et al.*, 2022; Pisacane and Halic 2017); (ii) a differential expression approach to define populations of siRNAs by comparing sRNAs in wildtype and certain mutant backgrounds (e.g., Gu *et al* 2009; Conine *et al.* 2010; Montgomery *et al.* 2008; Hua *et al.* 2021; Krzyszton and Kufel 2022); and (iii) computationally predicting siRNAs from sRNA sequence data (e.g., Montgomery *et al.* 2012, Wei *et al.* 2019, Rogers and Phillips 2020, Scheer *et al.* 2021). Differential expression and AGO immunoprecipitation have been very useful for parsing biogenesis mechanisms and linking individual Argonaute proteins to regulation of certain target genes via specific siRNA guides, although they do not provide a comprehensive view of all siRNA sequences in a transcriptome or their abundance. Computational pipelines have been very useful, but many of those developed for published studies are not being actively maintained nor were developed for general use across different datasets. We seek to overcome this limitation by developing and actively maintaining a user-friendly pipeline for identifying siRNAs within full sRNA-seq data sets.

SiRNA biogenesis varies among organisms, and we chose to identify siRNAs based on pattern rather than biogenesis mode. In other words, although most siRNAs are produced from dsRNA precursors in most organisms, the strand that is antisense to the target RNA is

what guides the Argonaute protein to that target. Consequently, we developed smalldisco, a *small*-interfering RNA *disco*very program, for parsing sRNA-seq data sets to identify siRNAs that are antisense to genomic features of interest to the user, such as protein-coding regions. While this approach is most obviously appropriate for *C. elegans*—where most siRNAs are direct RdRP products—we argue that it can provide useful information for other species, as we demonstrate below. Our goal was to develop a broadly applicable pipeline that is useful for addressing various biological questions. After identification, smalldisco can feed its discovered siRNAs into the Tailor program (Chou *et al.* 2015) to identify 3' non-templated nucleotide additions to these siRNA sequences and produce output in a user-friendly format as tables of siRNA targets, their tail modifications, and their abundances. While smalldisco does not identify miRNAs or piRNAs (programs available for those purposes include unitas [Gebert et al. 2017], miRDeep* [An et al. 2013], and piPipes [Han et al. 2015]), it does provide a wrapper option to identify 3' tailing modifications to miRNA and piRNA sequences from the same sRNA dataset through Tailor.

Methods

smalldisco users will need the following files: a folder of BAM files for their samples of sRNA-seq data; a reference genome (FASTA); and an annotated (GTF or GFF) file of the genome of interest. Note that the GTF/GFF file must in be the standard tab-separated, nine-column format (https://useast.ensembl.org/info/website/upload/gff.html). A summary of the pipeline is diagrammed in Figure 1. smalldisco can be run in two modes: `sirna` and `tail`.

The `sirna` mode implements smalldisco's siRNA discovery. First, the program creates a set of candidate genomic regions that could contain the siRNAs of interest. These are created from a specified annotated genomic feature, such as exons or 3' UTRs; overlapping features on the same strand are merged, while any regions of overlap across different strands are removed from the analysis. Second, reads are filtered to those mapping antisense to one of these candidate regions. Third, groups of overlapping reads are merged into a putative siRNA region if they fulfill two criteria: if there are more than *r* overlapping reads and if the putative siRNA would span *s* or more base pairs. Both parameters are set by the user, with default *r=10* and *s=10*. The result is a BED file of putative siRNA regions. The handling of siRNA sequences spanning a splice junction in an exon-exon boundary depends

on the alignment of reads. If there are antisense reads mapping to both exons, two putative siRNA regions will be reported, while if the reads map to only one of the exons, only one region in that exon will be reported. In the example datasets that we list below (Table S1), 0.01-1.5% of reads within a BAM file are sequences that span an exon-exon junction. Therefore, we recommend that users perform their read alignment with a splice-aware aligner such as HISAT2.

Another consideration is the detection of siRNA regions within transposable elements. If the user chooses to find siRNAs within exons, those would include transposase genes present in intact transposable elements. Repetitive sequences within intact or partial transposable elements could also be defined as the features producing candidate regions (but see the Readme for considerations). Given the complications associated with mapping repetitive element-derived reads to a genome, we recommend that the user map the sRNA reads to a FASTA file of consensus sequences of transposable elements that are derived from their genome of interest.

`tail` mode, the second and optional step of the pipeline, can quantify read tails for the sRNAs. A BED file of sRNA regions, which can be the list of putative siRNA regions generated by the previous step or any other predefined list, is used as input to Tailor (https://github.com/jhhung/Tailor). Tailor maps the reads to the reference genome and adds a tag at the end of each read's alignment record to indicate the length and sequence of a tail, or to indicate the absence of a tail, producing a tagged BAM file. Tailor implements a default of zero mismatches in the templated sRNA sequence before searching for non-templated tails. Tagged reads are then mapped to the provided sRNA regions file and the tail tags are converted to nucleotides. The output is a CSV file of sRNA targets and tails. In addition to "untailed" and the tail types in the table, one row per target will include a total for "all_tails." The sum of "untailed" and "all_tails" equals the total number of reads for that target.

We validated smalldisco using publicly available datasets published by Xu *et al.* (2018) (PRJNA447866), Claycomb *et al.* (2009) (PRJNA123485), Scheer *et al*. (2021) (PRJNA624174), and Davis *et al*. (2018) (PRJNA398116). Samples used for each analysis are listed in Table S1. To analyze these datasets, FASTQ files were obtained from NCBI's Gene Expression Omnibus, trimmed of 3' adapters using cutadapt, mapped to a reference genome using HISAT2, and run through smalldisco (both modes) with default parameters. For Xu *et al.* and Claycomb *et al.* datasets, the smalldisco siRNA output was normalized to

counts per million (CPM) and filtered to retain only siRNA-associated genes with ≥ 25 reads per million, the threshold used by these authors. To normalize the siRNA library as executed by Scheer *et al.*, we created a library by counting reads that map to exons, combined this library with the smalldisco siRNA output, and normalized to CPM. Finally, to benchmark smalldisco's performance against another sRNA discovery tool, segmentSeq (Hardcastle *et al.* 2011), we analyzed two samples from Davis *et al*. (2018), each with three replicates (SRR5931601, SRR5931602, SRR5931603, SRR5931604, SRR5931605, SRR5931606), and compared runtime, number of identified siRNA "loci," and resolution of siRNA region sizes.

Results and discussion

We validated the siRNA discovery function of smalldisco using publicly available datasets from two studies where *C. elegans* Argonaute-specific small RNAs were recovered by immunoprecipitation. Xu *et al.* (2018) immunoprecipitated WAGO-4 Argonaute and recovered siRNAs (present at or above 25 CPM) that corresponded to 4756 target genes. smalldisco analysis of their immunoprecipitated sRNA dataset identified siRNAs corresponding to 4706 target genes using a 25 CPM cut off, including ~93% of those identified by Xu *et al.* (4413) (Figure 2A). Claycomb *et al*. (2009) immunoprecipitated CSR-1 Argonaute and identified siRNAs corresponding to 4147 target genes present at or above 25 CPM and enriched ≥ two-fold in the IP compared to input RNA. smalldisco identified 4209 siRNA target genes in this dataset, including ~81% of those reported by Claycomb *et al.* (2009) (Figure 2B). For both datasets, some sequences were identified as siRNAs only by smalldisco or only by the authors, and we speculate these reflect differences in filtering criteria or data processing. Overall, our results demonstrate that smalldisco can accurately identify siRNAs with known Argonaute association.

We validated the siRNA counting function using publicly available datasets from Scheer *et al.* (2021). This study sequenced sRNAs from wildtype and three mutant strains of *Arabidopsis* and identified siRNAs within those four datasets. The authors provide a brief description of how they identified siRNA reads wherein their method uses logic similar to ours except that siRNAs were not required to map antisense to coding regions. We evaluated the four datasets with smalldisco and compared the output to the authors' results. For each genotype, the number of siRNA reads identified by smalldisco was similar to that

identified by the authors (Figure 3). Compared to the authors' analysis, smalldisco identified: 5% more siRNA reads in wildtype, the same number of reads in the *urt1-1* single mutant, 88% as many reads in the *xrn4-3* single mutant, and 85% as many reads in the *urt1-1 xrn4-3* double mutant. We note that our approach of identifying antisense reads uncovered a similar number of siRNAs as did the authors' approach of requiring matching sense and antisense reads.

One published protocol for identifying siRNAs and for which software is available is segmentSeq (Hardcastle *et al.* 2011). segmentSeq assumes all siRNAs are generated by Dicer endonuclease cleavage of double-stranded RNA, and therefore reads matching the sense strand should be present in the dataset. Although most *C. elegans* siRNAs are generated as direct single-strand RNA products of RNA-directed RNA polymerase and do not require Dicer activity, we were interested in comparing sRNA discovery by the two programs. We analyzed the Davis *et al.* (2018) dataset (see Methods) with each program and noted several major differences in output. First, smalldisco analyzed our sample input dataset in ~2.5 hr, whereas segmentSeq required ~2.5 weeks. Second, the two programs identified a common set of 6830 sRNA-producing genes and another smaller but substantial number of sRNA producing genes that are unique to one or the other program (1835 for segmentSeq, 3899 for smalldisco) (Figure 4A). Third, the two programs mapped sRNAs to the genome—and hence identified siRNA-associated genes—at different spatial scales. smalldisco identified an average siRNA region of 124.3 bp (± 7.4 bp SE) when run with the default "CDS" (exon) setting, whereas segmentSeq identified an average siRNA region of 6.7kb (± 68.9 bp SE) corresponding, in some cases, to more than one gene (Figure 4B, Figure S1). Although segmentSeq is a useful tool for siRNA identification, smalldisco gives finer grained results and is substantially faster.

In summary, smalldisco is a command-line tool for siRNA discovery and annotation from small RNA sequencing datasets. We expect that smalldisco will be useful in identifying, annotating, and quantifying the abundance of siRNAs in wildtype and mutant datasets derived from whole animals and dissected tissue. Inclusion of the Tailor program in the pipeline allows the user to identify and quantify 3' non-templated nucleotides that may be present on the siRNAs or any other small RNA in the dataset.

Data availability

smalldisco and supporting documentation is available on GitHub (https://github.com/ianvcaldas/smalldisco).

Acknowledgements
The authors thank Sarah Hall and two reviewers for helpful comments on the manuscript.

Funding
LHK was supported by a Syracuse University Graduate Fellowship. Funding was also provided by the National Institutes of Health (grant R03HD0091645) to EMM.

Conflicts of interest
The authors declare no conflicts of interest.

Literature cited

An J, Lai J, Lehman ML, Nelson CC. 2013. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res 41: 27-737.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Rev Genet 14:100-112.

Chaves DA, Dai H, Li L, Moresco JJ, Oh ME, Conte Jr D, Yates III JR, Mello CC, Gu W. 2021. The RNA phosphatase PIR-1 regulates endogenous small RNA pathways in *C. elegans*. Mol Cell 81:1-12.

Chou MT, Han BW, Hsiao CP, Zamore PD, Weng Z, Hung JH. 2015. Tailor: a computational framework for detecting non-templated tailing of small silencing RNAs. NAR 43(17):e109. doi: 10.1093/nar/gkv537.

Claycomb JM, Batista PJ, Pang KM, Gu W, Vasale JJ, van Wolfswinkel JC, Chaves DA, Shirayama M, Mitani S, Ketting RF, *et al*. 2009. The Argonaute CSR-1 and its 22G-RNA cofactors are required for homocentric chromosome segregation. Cell 139:123-134.

Conine CC, Batista PJ, Gu W, Claycomb JM, Chaves DA, Shirayama M, Mello CC. 2010. Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in *Caenorhabditis elegans*. PNAS USA 107:3588-3593.

Davis GM, Shikui T, Anderson JWT, Colson RN, Gunzburg MJ, Francisco MA, Ray D, Shrubsole SP, Sobotka JA, Seroussi U, Lao RX, Maity T, Su MZ, *et al.* 2018. The TRIM-NHL protein NHL-2 is a co-factor in the nuclear and somatic RNAi pathways in *C. elegans*. eLife 7:e35478. doi: 10.7554/eLife.35478.

de Albuquerque BFM, Placentino M, Ketting RF. 2015. Maternal piRNAs are essential for germline development following de novo establishment of endo-siRNAs in *Caenorhabditis elegans*. Dev Cell 34:228-256.

Galagali H, Kim JK. 2020. The multifaceted roles of microRNAs in differentiation. Curr Op Cell Biol 67:118-140.

Gebert D, Hewel C, Rosenkranz D. 2017. unitas: the universal tool for annotation of small RNAs. BMC Genet 18: 1-14.

Gartland S, Zeng B, and Marr MT III. 2022. The small RNA landscape is stable with age and resistant to loss of dFOXO signaling in Drosophila. PLoS One 17(11): e0273590.

Gu W, Shirayama M, Conte D Jr, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, Chen CC, Chaves DA, Duan S, Kasschau KD, Fahlgren N, Yates JR 3rd, Mitani S, Carrington JC, Mello CC. 2009. Mol Cell 36:231-244.

Han BW, Wang W, Zamore PD, Weng Z. 2015. piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics 31:593-595.

Hardcastle TJ, Kelly KA, Baulcombe DC. 2011. Identifying small interfering RNA loci from high-throughput sequencing data. Bioinformatics 28:457-463.

Holoch D, Moazed D. 2015. RNA-mediated epigenetic regulation of gene expression. Nature Rev Genet 16:71-84.

Hua X, Berkowitz ND, Willmann MR, Yu X, Lyons E, Gregory BD. 2021. Global analysis of RNA-dependent RNA polymerase-depend small RNAs reviews new substrates and functions for these proteins and SGS3 in Arabidopsis. Non-coding RNA 7, 28. doi:10.3390/ncrna7020028

Iwakawa H, Tomari Y. 2022. Life of RISC: formation, action, and degradation of RNA-induced silencing complex. Mol Cell 82: 31-43. doi: 10.1016/j.molcel.2021.11.026.

Ketting RF, Cochella L. 2021. Concepts and functions of small RNA pathways in *C. elegans*. Curr Top Dev Biol 144:45-89.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotech 37:907-915.

Krzyszton M and Kufel J. 2022. Analysis of mRNA-derived siRNAs in mutants of mRNA maturation and surveillance pathways in *Arabidopsis thaliana*. Sci Rep 12(1474).

Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, Chen S, Hannon GJ, Qi Y. 2008. Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell 133:116-127.

Montgomery TA, Yoo SJ, Fahlgren N, Gilbert SD, Howell MD, Sullivan CM, Alexander A, Nguyen G, Allen D, Ahn JH, Carrington JC. 2008. AGO1-miR173 complex initiates phased siRNA formation in plants. PNAS USA 105:20055-20062.

Montgomery TA, Rim YS, Zhang C, Dowen RH, Phillips CM, Fischer SEJ, Ruvkun G. 2012. PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog *henn-1*. PLoS Genet. 8:1-12.

Okamura K, Ladewig E, Zhou L, Lai EC. 2013. Functional small RNAs are generated from select miRNA hairpin loops in flies and mammals. Genes Dev 27:1-16.

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. Nature Rev Genet 20:89-108.

Pisacane P, Halic M. 2017. Tailing and degradation of Argonaute-bound small RNAs protect the genome from uncontrolled RNAi. Nat Commun 8:15332. doi: https://doi.org/10.1038/ncomms15332.

Rogers AR, Phillips CM. 2020. A small-RNA-mediated feedback loop maintains proper levels of 22G-RNAs in *C. elegans*. Cell Reports 33(3):1-13.

Scheer H, de Almeida C, Ferrier E, Simonnot Q, Poirier L, Pflieger D, Sement FM, Koechler S, Piermaria C, Krawczyk P, Mrocsek S, *et al.* 2021. The TUTase URT1 connects decapping activators and prevents the accumulation of excessively deadenylated mRNAs to avoid siRNA biogenesis. Nat Commun 12:1298. doi: 10.1038/s41467-021-21382-2.

Seroussi U, Lugowski A, Wadi L, Lao RX, Willis AR, Zhao W, Sundby AE, Charlesworth AG, Reinke AW, Claycomb JM. A comprehensive survey of *C. elegans* Argonaute proteins reveals organism-wide gene regulatory networks and functions. bioRxiv preprint doi: 10.1101/2922.08.08.502013.

Svendsen JM, Reed KJ, Vijayasarathy T, Montgomery BE, Tucci RM, Brown KC, Marks TN, Nguyen DAH, Phillips CM, Montgomery TA. 2019. Henn-1/HEN1 promotes germline immortality in *Caenorhabditis elegans*. Cell Rep 29:3187-3199.e4.

Wei Y, Li F, Zhang S, Zhang S, Zhang H, Sun R. 2019. Analysis of small RNA changes in different *Brassica napus* synthetic allopolyploids. PeerJ 7:e7621 https://doi.org/10.7717/peerj.7621

Xu R, Feng X, Chen X, Weng C, Yan Q, Xu T, Hong M, Guang S. 2018. A cytoplasmic Argonaute protein promotes the inheritance of RNAi. Cell Rep 23:2482-2494.

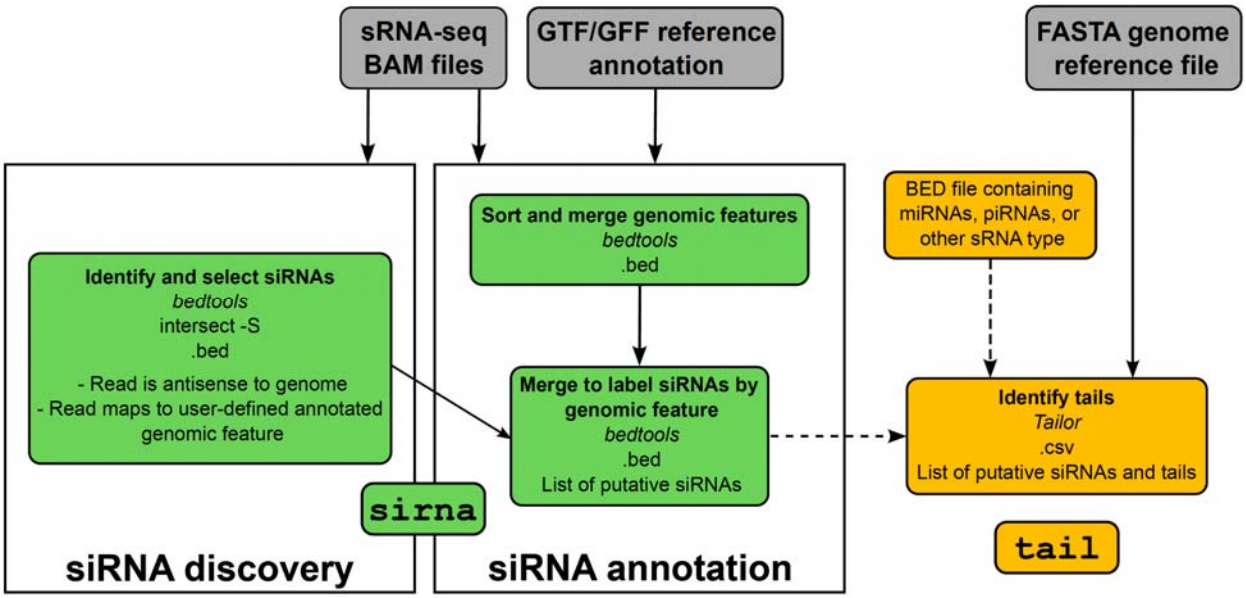


Figure 1 Overview of the smalldisco pipeline outlining the siRNA discovery and annotation workflows. Two analysis modes are available: `sirna` and `tail`. "Genomic feature" refers to a user-defined annotated genomic feature. Dashed arrows indicate only one of these inputs is necessary.

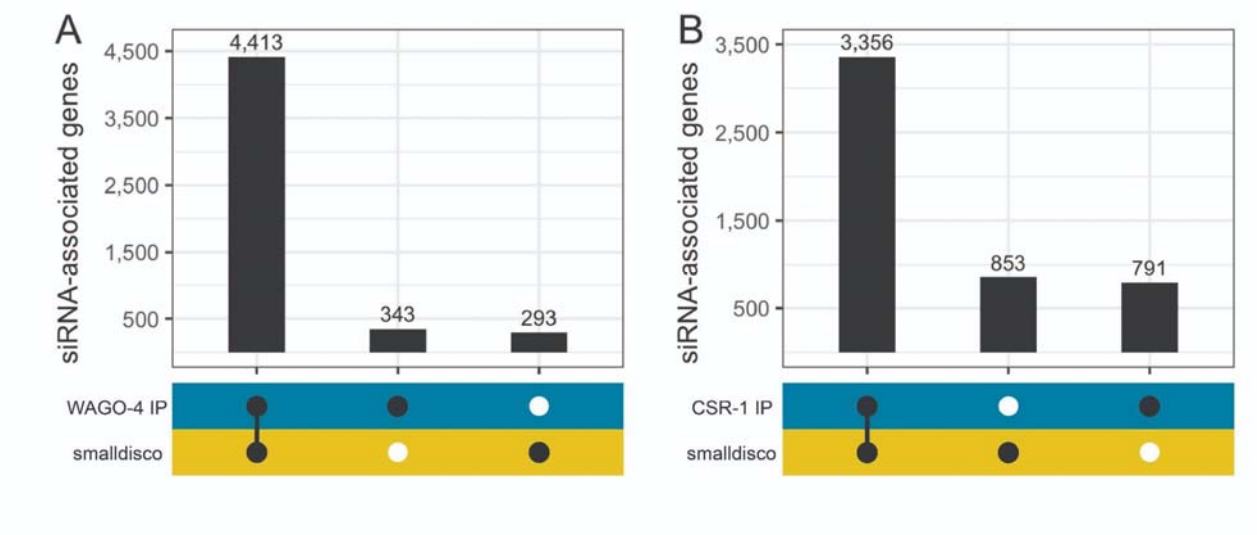


Figure 2 siRNA identification among sRNAs that co-immunoprecipitate with Argonaute proteins. Small RNAs enriched in (A) WAGO-4 immunoprecipitate (Xu *et al.* 2018) and (B) CSR-1 immunoprecipitate (Claycomb *et al.* 2009) were analyzed with smalldisco and the output was compared to the authors' reported findings. Plots indicate the number of siRNA target genes identified in the immunoprecipitated sRNA dataset by the authors' analysis (WAGO-4 IP, CSR-1 IP) and/or by smalldisco.

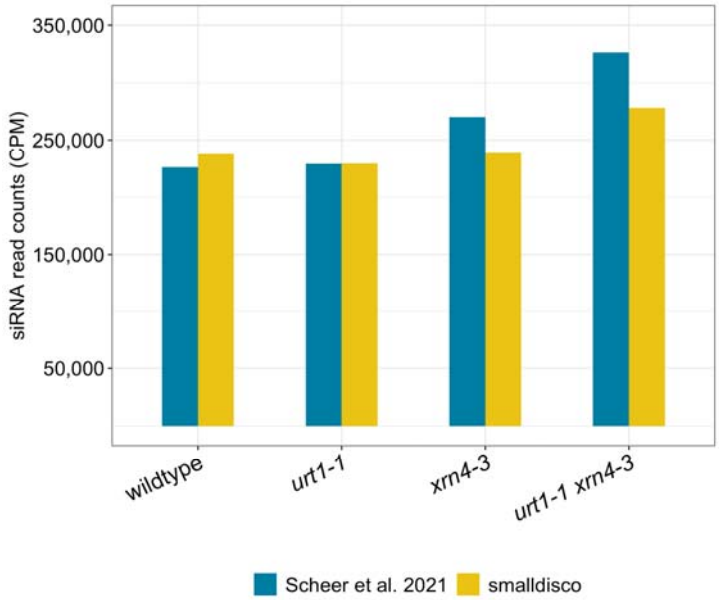


Figure 3 Identification of siRNA reads in small RNA sequencing datasets. SiRNA read counts are plotted as counts per million reads for small RNA sequencing datasets obtained for wildtype and three mutant genotypes. Blue bars indicate read counts reported by Scheer *et al.* (2021) and yellow bars indicate reads identified by smalldisco.

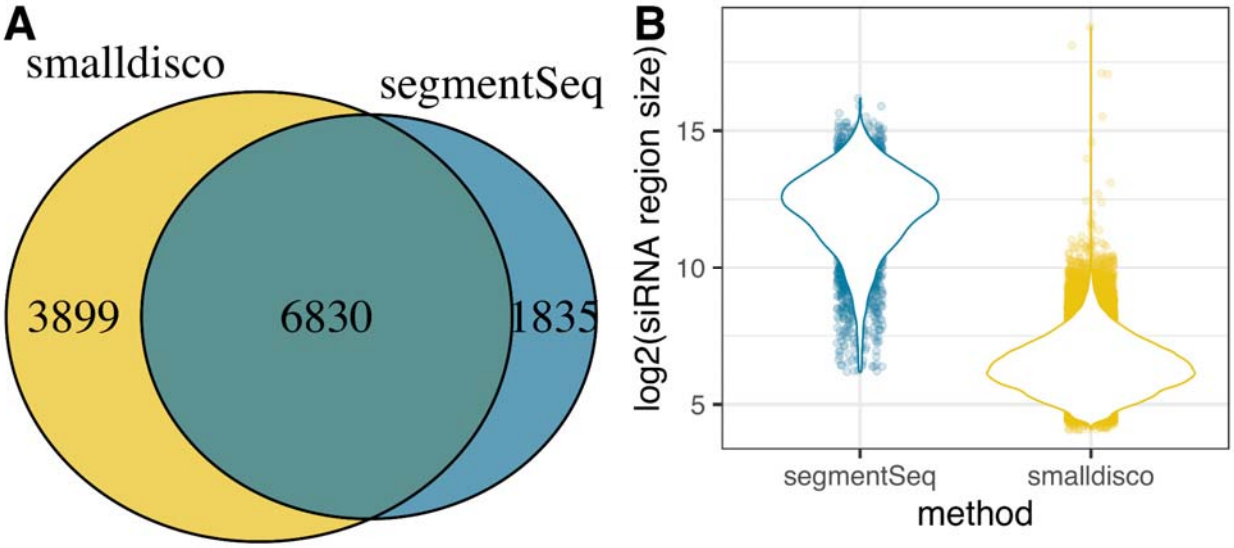


Figure 4 siRNA prediction comparison between smalldisco and segmentSeq. (A) Overlap between the number of sRNA target genes predicted using smalldisco and segmentSeq. Note that segmentSeq is not specific to siRNAs, but presumably most sRNAs derived from coding sequences are likely to be siRNAs. (B) The two methods differ substantially in the size of their predicted siRNA region, with smalldisco having an average size of 124.3 bp (±7.4 bp SE) and segmentSeq having an average size of 6.7kb (±68.9 bp SE).

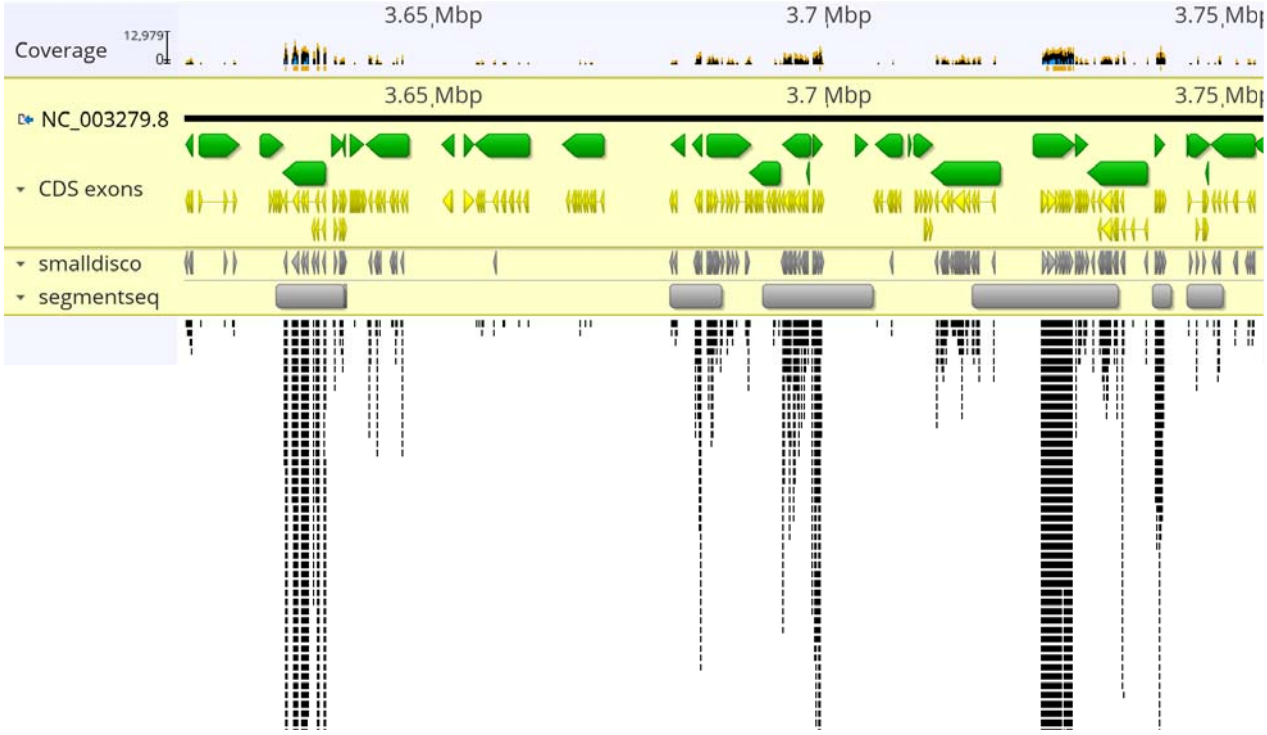


Figure S1 Examples of siRNAs identified by smalldisco and segmentSeq that map to one region of the genome. Image is a screen shot from the Geneious genome browser. Green bars indicate the positions of transcribed genes, with the pointed end indicating direction of transcription. Yellow bars indicate coding exons. Gray bars indicate siRNAs as identified by smalldisco (top) and segmentSeq (bottom). Stacked black lines represent siRNA-derived read pileups.

Table S1: List of NCBI accession numbers used in this study and frequency of siRNAs that map to exon-exon boundaries.

Tab 1

A	B	C	D	E
Paper	**BioProject**	**Run(s)**	**Sample description**	**Figure**
Xu et al. 2018	PRJNA447866	SRR6912753	3xFLAG::GFP::WAGO-4, 20°C	Figure 2A
Claycomb et al. 2009	PRJNA123485	SRR030720	mmunoprecipitate with the argonaute CSR-1; CSR-1 immunoprecipitat	Figure 2B
Scheer et al. 2021	PRJNA624174	SRR11521524	WT, replicate 1	Figure 3
		SRR11521526	urt1-1, replicate 1	
		SRR11521528	xrn4-3, replicate 1	
		SRR11521530	urt1-1 xrn4-3, replicate 1	
		SRR11521532	WT, replicate 2	
		SRR11521534	urt1-1, replicate 2	
		SRR11521536	xrn4-3, replicate 2	
		SRR11521538	urt1-1 xrn4-3, replicate 2	
Davis et al. 2018	PRJNA398116	SRR5931601	wildtype at 20C small RNA replicate 1	Figure 4
		SRR5931602	wildtype at 20C small RNA replicate 2	
		SRR5931603	wildtype at 20C small RNA replicate 3	
		SRR5931604	wildtype at 25C small RNA replicate 1	
		SRR5931605	wildtype at 25C small RNA replicate 2	
		SRR5931606	wildtype at 25C small RNA replicate 3	

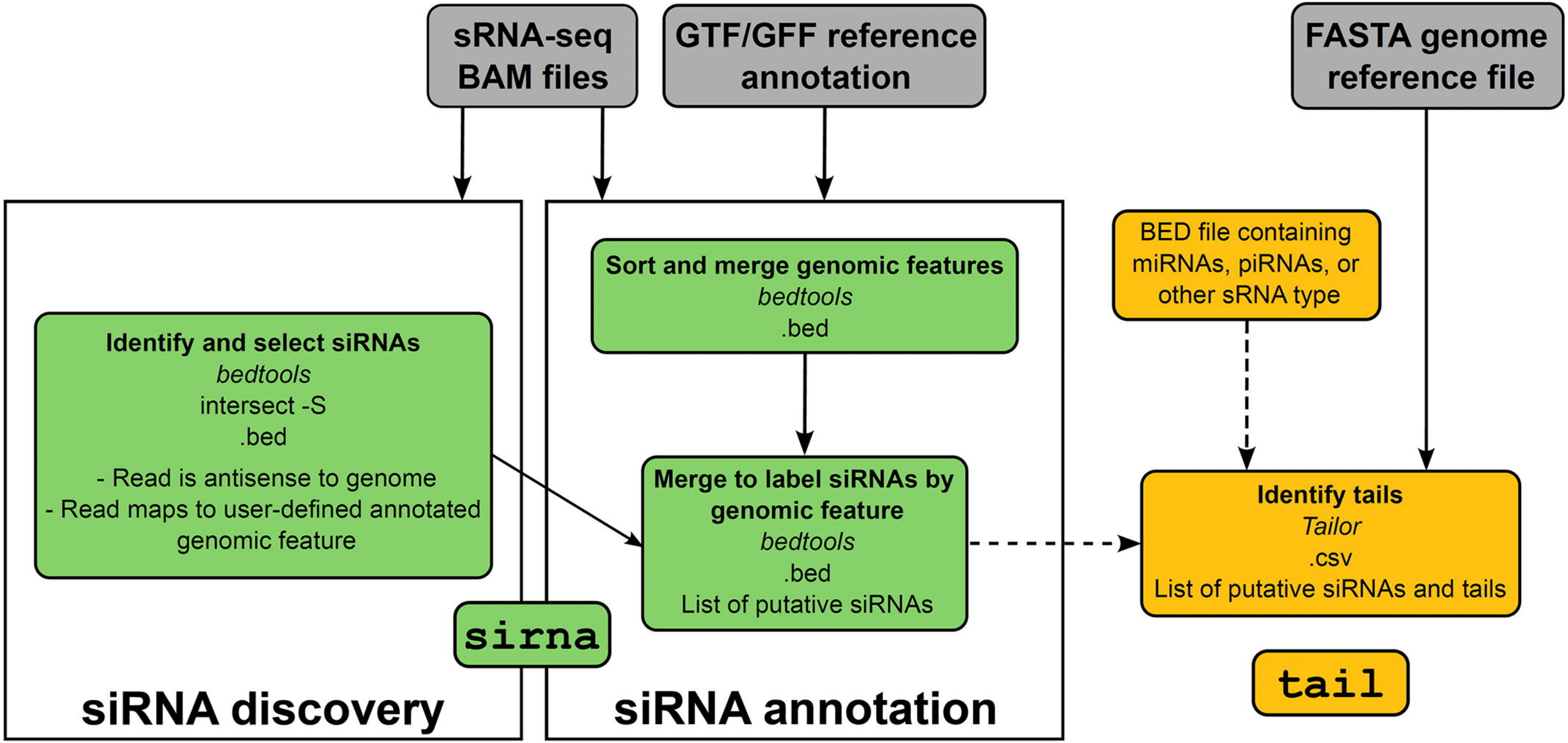
Paper	BioProject	Runs	Number of siRNA sequences identified by smalldisco that span exon-exon boundries	Total reads	Percentage of siRNAs that span exon-exon boundaries
Xu et al. (2018)	PRJNA447866	SRR6912753	25	12,805,988	0.0002
Claycomb et al. (2009)	PRJNA123485	SRR030720	264	2,277,181	0.0116
Scheer et al. (2021)	PRJNA624174	SRR11521524	761,474	107,145,803	0.7107
		SRR11521526	37,167	44,810,360	0.0829
		SRR11521528	261,853	49,815,405	0.5256
		SRR11521530	42,569	55,003,737	0.0774
		SRR11521532	43,695	45,909,295	0.0952
		SRR11521534	37,035	35,247,159	0.1051
		SRR11521536	225,826	51,870,377	0.4354
		SRR11521538	1,314,729	90,801,295	1.4479

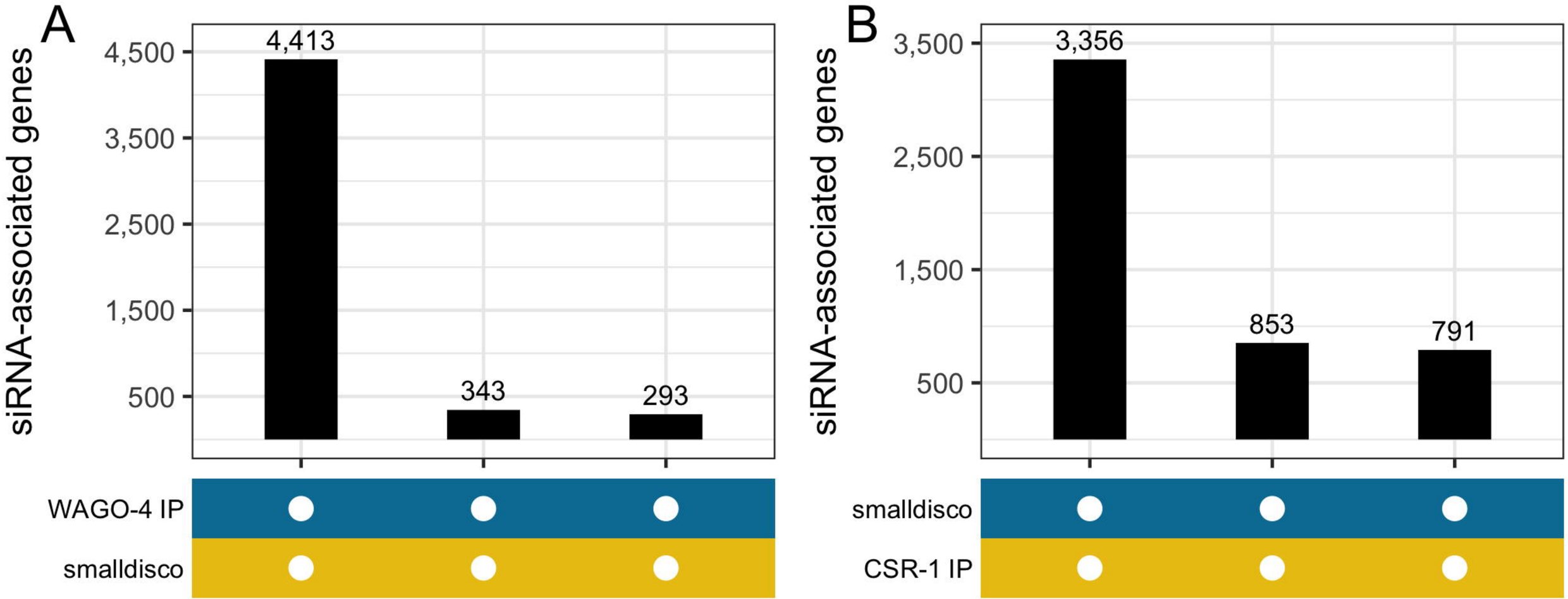
To obtain the number of siRNA sequences that span an exon-exon boundary (column D):

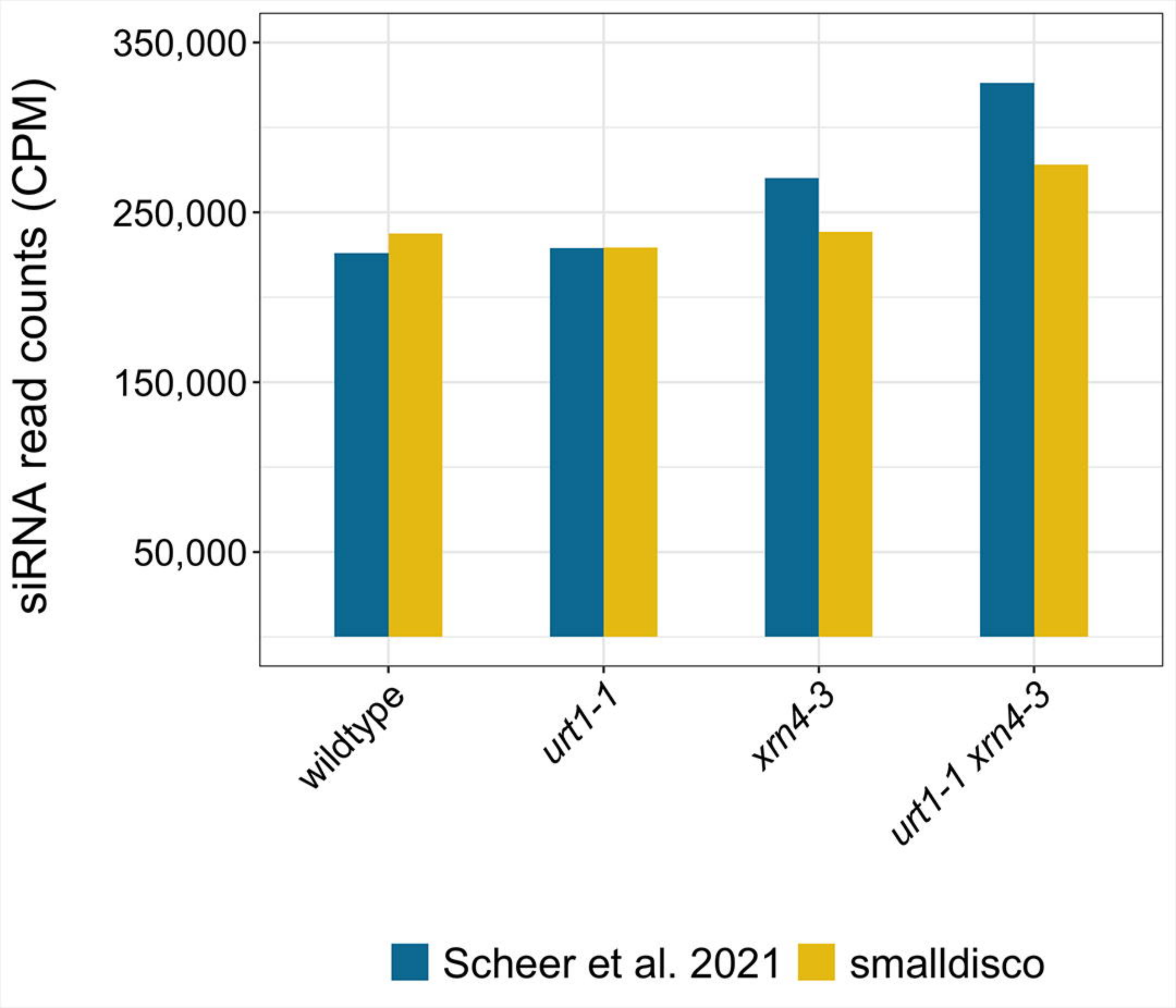
```
samtools view mapped.bam | awk '($6 ~ /N/)' | cut -f1 | wc -l
```

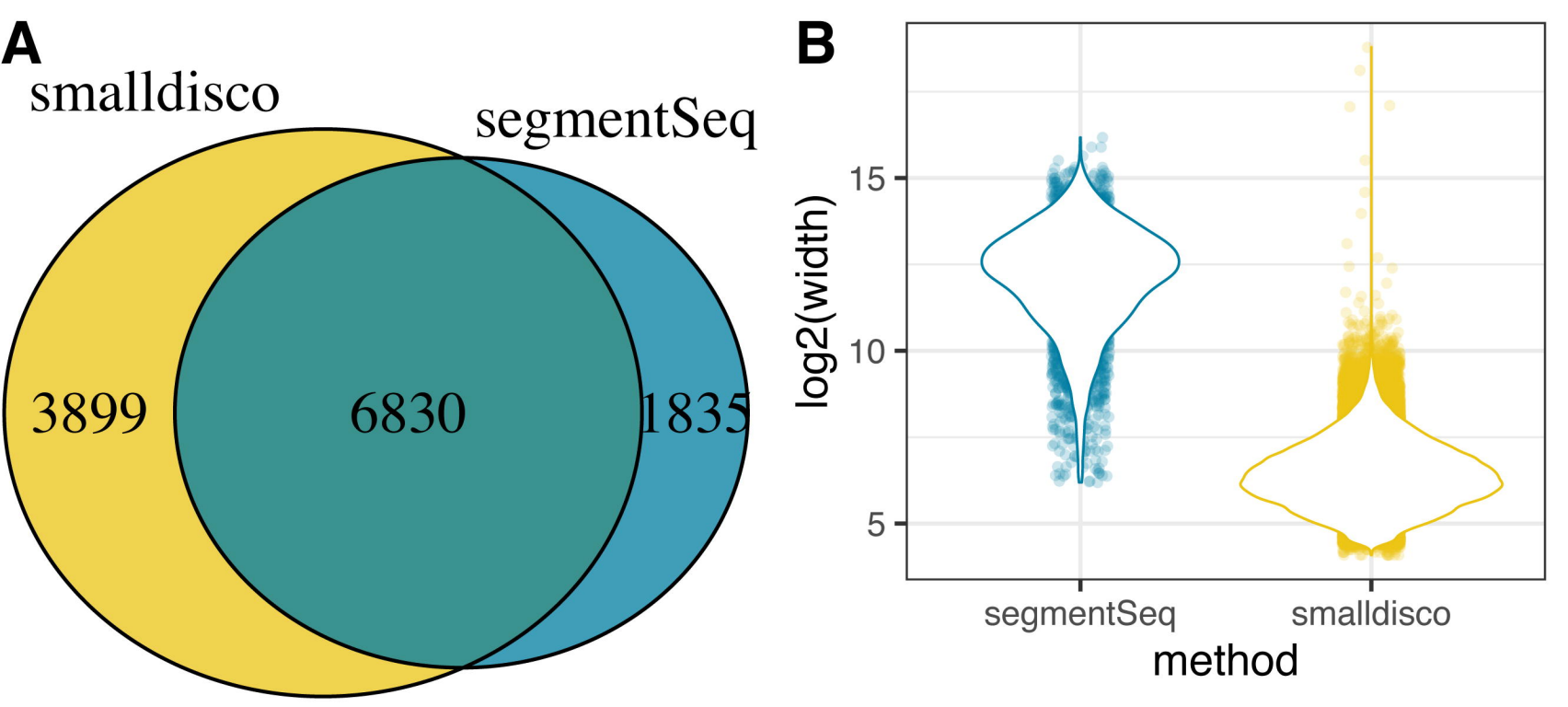
To obtain total sequences in bam file (column E):

```
samtools view mapped.bam | cut -f1 | wc -l
```


sRNA-seq BAM files
GTF/GFF reference annotation
FASTA genome reference file
Identify and select siRNAs
bedtools
intersect -S
.bed
- Read is antisense to genome
- Read maps to user-defined annotated genomic feature
sirna
siRNA discovery
Sort and merge genomic features
bedtools
.bed
Merge to label siRNAs by genomic feature
bedtools
.bed
List of putative siRNAs
siRNA annotation
BED file containing miRNAs, piRNAs, or other sRNA type
Identify tails
Tailor
.csv
List of putative siRNAs and tails
tail


A
siRNA-associated genes
4,500
3,500
2,500
1,500
500
4,413
343
293
WAGO-4 IP
smalldisco
B
siRNA-associated genes
3,500
2,500
1,500
500
3,356
853
791
smalldisco
CSR-1 IP


siRNA read counts (CPM)
350,000
250,000
150,000
50,000
wildtype
urt1-1
xrn4-3
urt1-1 xrn4-3
Scheer et al. 2021
smalldisco


A
smalldisco
segmentSeq
3899
6830
1835
B
log2(width)
15
10
5
segmentSeq
smalldisco
method